Intraclass Correlation Associated with Therapists: Estimates and Applications in Planning Psychotherapy Research

By: Scott A. Baldwin, David M. Murray, William R. Shadish, Sherri L. Pals, Jason M. Holland, Jonathan S. Abramowitz, Gerhard Andersson, David C. Atkins, Per Carlbring, Kathleen M. Carroll, Andrew Christensen, Kari M. Eddington, Anke Ehlers, Daniel J. Feaster, Ger P. J. Keijsers, Ellen Koch, Willem Kuyken, Alfred Lange, Tania Lincoln, Robert S. Stephens, Steven Taylor, Chris Trepka, Jeanne Watson

Baldwin, S.A., Murray, D.M., Shadish, W.R., Pals, S.L., Holland, J.M., ...(2011) Intraclass correlation associated with therapists: estimates and applications in planning psychotherapy research. *Cognitive Behaviour Therapy, 40 (1),* 15-33.doi: 10.1080/16506073.2010.520731

This is an Accepted Manuscript of an article published by Taylor & Francis Group in *Cognitive Behaviour Therapy* on 19 Feb 2011, available online at: http://www.tandfonline.com/10.1080/16506073.2010.520731

Abstract:

It is essential that outcome research permit clear conclusions to be drawn about the efficacy of interventions. The common practice of nesting therapists within conditions can pose important methodological challenges that affect interpretation, particularly if the study is not powered to account for the nested design. An obstacle to the optimal design of these studies is the lack of data about the intraclass correlation coefficient (ICC), which measures the statistical dependencies introduced by nesting. To begin the development of a public database of ICC estimates, the authors investigated ICCs for a variety outcomes reported in 20 psychotherapy outcome studies. The magnitude of the 495 ICC estimates varied widely across measures and studies. The authors provide recommendations regarding how to select and aggregate ICC estimates for power calculations and show how researchers can use ICC estimates to choose the number of patients and therapists that will optimize power. Attention to these recommendations will strengthen the validity of inferences drawn from psychotherapy studies that nest therapists within conditions.

Keywords: statistical dependence | therapist effects | intraclass correlation | power | psychotherapy research

Article:

Many psychotherapy outcome studies use more than one therapist to administer the intervention; in these studies, it is common to have patients nested within therapists and therapists nested within conditions. This nesting creates an opportunity to study therapist effects (Wampold, 2001), but it generates statistical dependencies that can lead to erroneous conclusions about both treatment outcomes and therapist effects (Crits-Christoph & Mintz, 1991; Wampold & Serlin, 2000).

Optimal design of these studies requires a good estimate of the dependence expected among observations of patients who have the same therapist, indexed by the intraclass correlation coefficient (ICC). Unfortunately, because most psychotherapy outcome studies have not considered this issue, few have reported ICCs for therapists. To address this problem, our research group has begun building a database of therapist ICCs for outcomes commonly used in psychotherapy research. ICC databases in public health and education have helped researchers design studies in those disciplines (cf. Donner & Klar, 2000; Gulliford, Ukoumunne, & Chinn, 1999; Hedges & Hedberg, 2007; Murray & Blitstein, 2003; Murray et al., 1994; Murray, Varnell, & Blitstein, 2004; Verma & Le, 1996). The purposes of this study are to (1) review the statistical effects of nesting and explain essential concepts, (2) report therapist ICCs from recent psychotherapy outcome studies, (3) provide guidelines for selecting and aggregating ICC estimates, and (4) show how to use ICC estimates to design new studies.

Nested designs in psychotherapy research

Nested designs in psychotherapy research can take many forms. Patients may be seen in groups or as individuals. Randomization may occur at the level of the patient, the therapist, or both, or there may be no randomization. Nesting may occur in all conditions or in some conditions but not others (e.g. a wait-list condition). This study focuses on the statistical issues that exist in all studies in which patients are nested within therapists and therapists are nested within at least one condition. Previous reports discuss the design and analysis of nested designs generally (e.g. Cornfield, 1978; Murray et al., 2004; Pals et al., 2008; Roberts, 1999; Roberts & Roberts, 2005; Schnurr, Friedman, Lavori, & Hsieh, 2001; Zucker, 1990) and in psychotherapy research (e.g. Baldwin, Murray, & Shadish, 2005; Crits-Christoph & Mintz, 1991; Martindale, 1978; Wampold & Serlin, 2000). We add to this work by providing therapist ICC estimates and illustrating their use in power calculations for psychotherapy research.

A key concept in these nested psychotherapy designs is *statistical dependence* associated with therapist. Observations are dependent if they are correlated, and in any nested design, we expect some level of correlation (Cornfield, 1978; Kish, 1965; Zucker, 1990). Where patients select their therapist or are assigned to a therapist based on a nonrandom assignment rule, their observations may be correlated even before therapy begins as a result of shared selection factors or prior exposures. Once patients are assigned to therapists, their observations may become correlated over time through mutual interaction and common exposures, including exposure to

the same therapist. Whatever the origin, the degree of within-therapist correlation is indexed with an ICC (Kenny, Mannetti, Peirro, Livi, & Kashy, 2002).

Like other correlations, ICCs can be positive or negative. Positive ICCs may reflect differential effectiveness among therapists as a result of their skill level in developing a working alliance with patients (cf. Baldwin, Wampold, & Imel, 2007), their general competence, their adherence to treatment protocols, or any variable that differs among therapists. A meta-analysis of ICCs from 15 psychotherapy clinical trials found that ICCs varied widely (range = 0–.729), with a mean of about .08 (Crits-Christoph et al.,1991). Similar results have been found in clinical trial (Elkin, Falconnier, Martinovich, & Mahoney, 2006; Kim, Wampold, & Bolt,2006) and clinical practice (Baldwin et al., 2007; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Okiishi, Lambert, Nielsen, & Ogles, 2003; Wampold & Brown, 2005) data. Ignoring positive ICCs increases the rate of Type I errors, that is, concluding that a treatment is effective when it is not (Crits-Christoph & Mintz, 1991; Cornfield, 1978; Pals et al., 2008; Wampold & Serlin, 2000; Zucker, 1990).

Negative ICCs could occur if patients responded to the same therapist differently. Whereas most patients enter treatment functioning relatively poorly, some leave treatment functioning well, some do not change, and some deteriorate (Bergin, 1966). If this increased variability occurs more within therapists than between therapists, a negative ICC could result. Negative ICCs could occur if there is a competition among patients within a therapist (e.g. competition for attention from a therapist in a group therapy setting). Negative ICCs could occur if there is an unequal distribution of resources (e.g. a therapist getting burned out toward the end of a study). Ignoring negative ICCs reduces the Type I error rate and so reduces power, that is, mistakenly concluding that a treatment is ineffective (Kenny et al., 2002; Murray, Hannan, & Baker, 1996; Swallow & Monahan, 1984).

Regardless of whether the population ICC is positive or negative, large or small, ICCs can be estimated as positive or negative, large or small, as a result of sampling error. If the population value of the ICC is close to zero, ICCs will be estimated as negative about half the time. If the sample size is small, there will be less precision in the estimates, so that some estimates may be quite large or small relative to the population value, whether positive or negative. The best way to obtain an accurate estimate of the population value is to estimate the ICC from studies involving many therapists and patients or by pooling estimates from several smaller studies.

Expected within-therapist dependence must be addressed when the study is planned in order to ensure adequate power, and to do so, investigators need good estimates of ICCs appropriate to their study. In the next sections, we report 495 ICC estimates drawn from 20 recent psychotherapy studies. In addition, we provide power formulas and illustrate how to use the ICCs from our database to plan future studies.

Intraclass correlation database

Studies

We identified potential studies in two ways. First, we performed manual searches for the years 2003–2004 of journals that regularly publish psychotherapy research (*Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology, Behavior Therapy, Behaviour Research and Therapy, Archives of General Psychiatry, American Journal of Psychiatry, Psychotherapy Research,* and *Cognitive Therapy Research*). Recent issues were used to increase the likelihood that authors would be familiar with and interested in using their data in this study. We sampled broadly with respect to treatment type and design so that the database would be as useful as possible. Potential studies had to include at least one condition that involved an intervention aimed at reducing an emotional or behavioral problem. Therapy had to be delivered to individuals and not to groups. In addition, participants in the intervention conditions had to interact with a therapist. Given the increasing use of Internet-based treatments, we included them if the treatment involved interaction with a therapist (e.g. supplemental phone calls). Finally, studies had to include at least two therapists per condition and each therapist had to see at least two patients so that we could distinguish between therapist and patient variability.

The manual search produced 38 potential studies. We wrote to the corresponding authors and asked them to perform several analyses that would allow us to compute ICCs (see later) and provide us with the output. We also invited them to join us as coauthors on the resulting manuscript. Sixteen authors agreed to participate in the study. Of the 22 authors who did not participate, the majority indicated time constraints as the reason. One study in our database was published after 2004 (Carlbring et al., 2006). The original study we contacted the corresponding author about did not meet our inclusion criteria; however, the author suggested that we use data from a newer study instead. Because we wanted to calculate as many ICC estimates as possible, we included Carlbring et al. (2006) in our database.

Our second method for identifying studies was to locate published ICC estimates from psychotherapy outcome studies. We performed an electronic literature search using the terms *intraclass correlation*, *therapist variability*, or *therapist effect*. Additionally, we reviewed the reference section of articles on therapist effects. We located four new studies that published ICC estimates (Baldwin et al., 2007; Dinger, Strack, Leichsenring, Wilmers, & Schauenburg, 2008; Kim et al., 2006; Wampold & Brown, 2005).

Calculation of ICCs

We calculated ICCs from an analysis of variance (ANOVA) source table; therapist was included in the model as a fixed effect. The information from the source table was inserted into the following formula:

$$\hat{\rho} = \frac{MS_{therapist} - MS_{error}}{MS_{therapist} + (\bar{n} - 1)MS_{error}} \qquad (1)$$

where $MS_{therapist}$ is the mean square for therapist, MS_{error} is the mean square for patient, and $\bar{m}$ is the average number of patients per therapist. Because the number of patients per therapist varied within studies, we used the harmonic mean. This formula is appropriate both for continuous (Snedecor & Cochran, 1989) and dichotomous (Fleiss, Levin, & Paik, 2003) outcomes. The ICC will be positive when $MS_{therapist} > MS_{error}$, negative when $MS_{therapist} < MS_{error}$, and zero only if $MS_{therapist} = MS_{error}$. Donner (1986) noted that this ANOVA estimator is consistent but slightly biased, although the degree of bias is usually ignorable.

We asked each study author to calculate a one-way ANOVA with therapist as the independent variable for each outcome variable separately at pretest and at posttest. We also requested an analysis of covariance (ANCOVA) with therapist as the independent variable and the pretest value of the dependent variable as the covariate. To control for differences among treatment types, all analyses were done separately for each treatment condition. Each author provided us with the $MS_{therapist}$ and MS_{error} from each ANOVA and ANCOVA, and the information needed to determine $\bar{m}$. Thus, for each outcome variable and treatment type, we calculated a pretest ICC $(\hat{\rho}_{pre})$, a posttest-only ICC $(\hat{\rho}_{post})$, and a posttest adjusted for pretest ICC $(\hat{\rho}'_{post})$.

It is also possible to calculate ICCs from the output of a mixed-model ANOVA or ANCOVA. Such programs generally do not provide mean squares and instead provide estimates of components of variance. We did not use this approach because the default for most such programs is to constrain all estimates to be nonnegative. As we discuss later, this approach prevents calculation of negative ICCs, with a number of unintended consequences.

Study characteristics

Tables 1 and 2 provide descriptive information about the 20 studies. Most treatments were behavioral or cognitive behavioral. Sixteen studies used a treatment manual. The number of sessions was fixed in some studies and allowed to vary in others. The number of sessions ranged from 1 to 22.9 (*Mdn* = 10.6). The number of therapists (*k*) ranged from 2 to 581 (*Mdn* = 5). The number of patients per therapist (*m*) ranged from 2.2 to 51.1 (*Mdn* = 4.9). Table 2 provides the study-level averages for *k* and *m*. Eight studies used full-time clinicians, one used PhD-level clinical researchers, four used clinicians-in-training, two used both full-time clinicians and PhD-level clinical researchers, three used clinical researchers and clinicians in training, and two used full-time clinicians and clinicians in training.

Table 1 Descriptive information for each study regarding sample size and the intervention delivered

Study	***N***	**Treated problem**	**Treatment type (no. sessions)**	**Manual**
			Efficacy studies	

1. Abramowitz, Foa, & Franklin (2003)	40	OCD	1. ERP: intensive (15)	Yes
			2. ERP: twice weekly (15)	
2. Carlbring et al. (2006)	30	Panic disorder	Internet-based CBT with supplemental phone calls (10)	Yes
3. Carroll et al. (2004)	104	Cocaine dependence	1. CBT+medication (12)	Yes
			2. CBT+placebo (12)	
			3. IPT+medication (12)	
			4. IPT+Placebo (12)	
4. Christensen et al. (2004)	134	Marital distress	1. IBCT (22.9)	Yes
			2. TBCT (22.9)	
5. Ehlers et al. (2003)	28	PTSD	CT (11.4)	Yes
6. Kim et al. (2006)	86	Depression	1. CBT (16.2)	Yes
			2. IPT (16.2)	
7. Koch, Spates, & Himle (2004)	40	Small animal phobia	1. Behavioral exposure (1)	Yes
			2. Cognitive behavioral exposure (1)	
8. Lange et al. (2003)	69	Posttraumatic stress	Interapy (10)	Yes
9. Marijuana Treatment Project Research Group (2004)	276	Cannabis dependence	1. MET (2) 2. MET, CBT, and case management (9)	Yes
10. Szapocznik et al. (2004)	129	HIV-positive African Americans: distress, hassles, support	1. SET (12.15) 2. PCA (6.78)	Yes
11. Taylor et al. (2003)	60	PTSD	1. EMDR (8)	Yes
			2. Exposure (8)	
			3. Relaxation (8)	

12. van Minnen, Hoogduin, Keijsers, Hellenbrand, & Hendriks (2003)	15	Trichotillomania	BT (6)	Yes
13. Watson, Gordon, Stermac, Kalogerakos, & Steckley (2003)	66	Depression	1. CBT (16) 2. PET (16)	Yes
			Effectiveness studies	
14. Baldwin et al. (2007)	331	Mixed	TAU (7.32)	No
15. Dinger, Strack, Leichsenring, Wilmers, & Schauenburg (2008)	2554	Mixed (inpatients)	Inpatient TAU[a]	No
16. Kuyken (2004)	105	Depression	CT (14.11)	Yes
17. Lincoln et al. (2003)	147	Social phobia	CBT (40)	No[b]
18. Merrill, Tolbert, & Wade (2003)	186	Depression	CT (7.8)	Yes
19. Trepka, Rees, Shapiro, Hardy, & Barkham (2004)	30	Depression	CT (15.52)	Yes
20. Wampold & Brown (2005)	6146	Mixed	TAU (10.63)	No

Note. CBT, cognitive behavior therapy; CT, cognitive therapy; EMDR, eye movement desensitization and reprocessing; ERP, exposure and response prevention; IPT, interpersonal therapy; IBCT, integrative behavioral couples therapy; MET, motivational enhancement therapy; OCD, obsessive–compulsive disorder; PCA, person-centered approach; PET, process-experiential therapy; PTSD, posttraumatic stress disorder; SET, structural ecosystems therapy; TBCT, traditional behavioral couples therapy; TAU, treatment as usual. [a] Specific number of sessions during the hospital stay was not available. [b] Although there was no treatment manual for this study, the treatment was structured: in vivo exposure and cognitive restructuring.

Table 2 Descriptive information for each study regarding therapists

Study	***k***	***m***	**Type of therapists**	**Therapist training**	**Therapist supervised**
		Efficacy studies			

1. Abramowitz et al. (2003)	5	2.87	FTC and CR	Yes	Yes
2. Carlbring et al. (2006)	3	7.43	CR and NC/CT	Yes	Yes
3. Carroll et al. (2004)	5.25	3.86	FTC and CR	Yes	Yes
4. Christensen et al. (2004)	7	8.39	FTC	Yes	Yes
5. Ehlers et al. (2003)	3	8.31	FTC	Yes	Yes
6. Kim et al. (2006)	17	5.00	FTC	Yes	Yes
7. Koch et al. (2004)	4	3.62	NC/CT	Yes	Yes
8. Lange et al. (2003)	18	2.92	NC/CT	Yes	Yes
9. Marijuana Treatment Project Research Group (2004)	12	7.04	FTC	Yes	Yes
10. Szapocznik et al. (2004)	3	18.40	CR and NC/CT	Yes	Yes
11. Taylor et al. (2003)	2	7.75	CR	Yes	Yes
12. van Minnen et al. (2003)	5	2.51	NC/CT	Yes	Yes
13. Watson et al. (2003)	7.5	4.25	CR and NC/CT	Yes	Yes
		Effectiveness studies			
14. Baldwin et al. (2007)	80	4.1	FTC and NC/CT	No	No[a]
15. Dinger et al. (2008)	50	51.1	FTC and NC/CT	No	No[a]
16. Kuyken (2004)	20	3.32	FTC	Yes	Yes
17. Lincoln et al. (2003)	9.5	2.92	NC/CT	Yes	Yes

18. Merrill et al. (2003)	8	17.19	FTC	Yes	Yes
19. Trepka et al. (2004)	6	4.18	FTC	Yes	Yes
20. Wampold & Brown (2005)	581	9.68	FTC	No	No

Note. k, study-level mean number of therapists contributing to any given intraclass correlation; m, study-level mean number of patients per therapist contributing to any given intraclass correlation; CR, clinical researchers; FTC, full-time clinicians; NC/CT, nonclinicians/clinicians-in-training; [a] Although there was supervision for the therapists in training, the supervision was not explicitly a part of the study.

Intraclass correlation estimates

We were able to compute $N = 152$ estimates of $\hat{\rho}_{pre}$, $N = 170$ estimates of $\hat{\rho}_{post}$, and $N = 164$ estimates of $\hat{\rho}^{I}_{post}$. There were fewer estimates of $\hat{\rho}_{pre}$ and $\hat{\rho}^{I}_{post}$ than $\hat{\rho}_{post}$ because several outcome variables involved behavior during treatment or were posttest-only variables, making baseline values or adjusting for baseline values impossible. Additionally, we located $N = 1$ published estimates of $\hat{\rho}_{post}$ and $N = 8$ published estimates of $\hat{\rho}^{I}_{post}$. The number of estimates per study ranged from 1 to 36 ($Mdn = 6$) for $\hat{\rho}_{pre}$, 1 to 42 ($Mdn = 6$) for $\hat{\rho}_{post}$, and 1 to 36 for $\hat{\rho}^{I}_{post}$ ($Mdn = 5.5$).[1]

The distributions for $\hat{\rho}_{pre}$, $\hat{\rho}_{post}$, and $\hat{\rho}^{I}_{post}$ were symmetric and quite similar. The estimates for $\hat{\rho}_{pre}$ ranged from − .475 to .579 ($Q1 = -0.099$, $Q2 = -0.014$, $Q3 = 0.063$) and 54.6% were negative. The estimates for $\hat{\rho}_{post}$ ranged from − .345 to .532 ($Q1 = -0.113$, $Q2 = -0.026$, $Q3 = 0.046$) and 63.7% were negative. The estimates for $\hat{\rho}^{I}_{post}$ ranged from − .343 to .45 ($Q1 = -0.104$, $Q2 = -0.018$, $Q3 = 0.079$) and 57% were negative.

Applications

Selecting and combining estimates

Researchers can use these ICC estimates to plan future psychotherapy studies involving therapists nested within conditions. When multiple ICC estimates are available, the precision of the power calculations can be increased by meta-analytically combining them (Blitstein, Hannan, Murray, & Shadish, 2005). Methodologists typically recommend that researchers only aggregate estimates from studies that are closely matched to their planned study (Blitstein et al., 2005; Murray, 1998). For example, Blitstein et al. recommend that researchers only combine estimates

from studies that had outcomes, research designs, and statistical analyses similar to those planned for the new study. That recommendation is based on the authors' experience that ICCs vary appreciably as a function of those variables.

We followed Blitstein et al.'s recommendation and combined estimates that came from the same measure, research design, and statistical model. Specifically, we combined ICC estimates for the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the most common measure in our database (N = 12; one study used the BDI-II; Beck, Steer, & Brown, 1996), and we combined estimates separately for efficacy (N = 8) and effectiveness (N = 4) studies, as some have speculated that ICCs from efficacy studies may be smaller than ICCs from effectiveness studies because of the higher levels of control and standardization in efficacy studies (Elkin et al., 2006). We used a Q test to determine whether there was significant heterogeneity among the ICCs and I^2 to determine the proportion of the total variation in the ICCs that is due to heterogeneity rather than chance (Higgins & Thompson, 2002). We report only the results from the adjusted posttest analyses because adjusted analyses are more common than unadjusted analyses, although in general results from the unadjusted analyses were similar.

For the BDI data, the random effects mean ICC for the effectiveness studies was $\bar{\rho}^{I}$ .049 and not statistically significant (p = .335). There was no heterogeneity among the estimates, I^2 = 0, $Q(3)$ = 0.981, p = .821, suggesting that the mean ICC is a good estimate for power calculations. For the efficacy studies, the mean ICC was $\bar{\rho}^{I}$ − .073 and not statistically significant (p = .21). Heterogeneity was large and statistically significant, I^2 = 75.49, $Q(7)$ = 24.478, p < .001. Thus, the mean ICC from the efficacy studies would not be a good estimate for power calculations. This may be a consequence of the fact that three of the four effectiveness studies involved cognitive therapy for the treatment of depression, whereas the efficacy studies were more heterogeneous and involved a variety of treatments aimed at a variety of problems. In situations like this, the investigator should choose the individual study that most closely matches the planned study and use the ICC estimate from that study to plan the new study.

There is one important caveat to remember regarding the selection of ICC estimates for power calculations. When the population ICC is close to zero, the probability that the ICC will be estimated as negative is high; further, the probability increases as the number of patients per therapist decreases. When the negative value is likely the result of sampling error, it would be imprudent to assume that the ICC in the new study will also be negative. If the investigator makes that assumption but the ICC in the new study proves to be positive, the new study may be substantially underpowered. On the other hand, if the investigator takes a conservative approach and uses a positive ICC in the power calculations, he or she will have some insurance against an underpowered study and will have extra power if the ICC in the new study turns out to be smaller than the value used in the power calculations. For this reason, when the best ICC estimate is negative and based on a large number of therapists, we recommend that researchers

use a small but positive ICC (e.g. .01) in their sample size calculations. This a conservative approach, but it will ensure that sample size will be sufficient even if the ICC proves to be small but positive. When the estimate is negative and based on a small number of therapists, we recommend that researchers use a larger positive ICC (e.g. .025 or .05) or a range of values (e.g. 0, .025, .05) to see how sample size requirements change as the ICC varies. The goal is to avoid underestimating the ICC, which would underpower the new study, but also to avoid overestimating the ICC, which would lead to a new study that was larger than it needed to be. This situation is particularly challenging for psychotherapy research, where most of the existing studies are small.

Example power analyses

We now present detectable difference formulas that can be used to plan new psychotherapy trials where therapists will be nested within conditions. We illustrate the use of the formulas for two common designs: a trial comparing two treatments involving therapists and a trial comparing a treatment involving therapists with one that does not.

Detectable difference: treatment versus treatment

The formula for this detectable difference is adapted from Murray (1998):

$$\Delta = \sqrt{\sigma_y^2\left(\frac{(1+(m_1-1)\rho_1)}{N_1}+\frac{(1+(m_2-1)\rho_2}{N_2}\right.}$$

where Δ is the detectable difference between condition means (i.e. the treatment effect); σ_y^2 is the variance of the dependent variable; m_1 and m_2 are the number of patients per therapists for Conditions 1 and 2; ρ_1 and ρ_2 are the ICCs for Conditions 1 and 2; N_1 and N_2 are the number of participants in Conditions 1 and 2; $t_{critical:\alpha/2}$ is the critical value for t needed to ensure the Type I error rate is α given a two-tailed test and available degrees of freedom; and $t_{critical:\beta}$ is the critical value for t needed to ensure the Type II error rate is β.

In the approach we present, the degrees of freedom for $t_{critical:\alpha/2}$ and $t_{critical:\beta}$ are k_1+k_2-2, where k_1 and k_2 are the number of therapists in Conditions 1 and 2, respectively. However, there may be situations (e.g. highly unbalanced designs, small ICCs) in which the Satterthwaite (1946) approximation may be needed (Roberts & Roberts, 2005). When using our approach, we recommend that k_1 and k_2 be equal, because this protects against inflated Type I error rates when there is heteroscedasticity in the therapist variance component (Gail, Mark, Carroll, Green, & Pee, 1996). Roberts & Roberts (2005) argue that if $\rho_1 \neq \rho_2$ and/or $m_1 \neq m_2$, power will be maximized for a given sample size by allocating more patients to the condition with the greatest variance inflation. However, their optimal allocation ratio assumes a fixed total sample size and static values for m_1 and m_2, requiring imbalance in k_1 and k_2. In light of Gail et al.'s (1996) findings, we recommend that researchers compare various combinations of m_1 and m_2—

ensuring that $k_1 = k_2$—to find the most powerful combination of values given the study's hypotheses and budgetary constraints.

Detectable difference: treatment versus comparison condition

Equation 2 requires a slight modification for designs comparing a treatment involving therapists (Condition 1) with one that does not (Condition 2). Because the patients in the comparison condition do not interact with each other or with a common therapist, observations in the comparison condition are independent, $\rho_2 = 0$ and Equation 2 reduces to:

A reasonable estimate for the degrees of freedom for this design is $k_1 + N_2 - 2$, although there may be situations in which the Satterthwaite approximation may be needed.

$$\Delta = \sqrt{\sigma_y^2\left(\frac{(1+(m_1-1)\rho_1)}{N_1}+\frac{1}{N_2}\right)(t_{critical:\alpha/2}+t_{critical:\beta})^2} \qquad (3)$$

Because therapists are involved in only one condition, imbalance in the number of therapists per condition is unavoidable. When the ICC due to therapist is greater than zero, power will be maximized by allocating more patients to the condition involving therapists. The optimal allocation ratio (R) is (Roberts & Roberts, 2005):

$$R = \sqrt{(1+(m_1-1)\rho_1)} \qquad (4)$$

where ρ_1 is the ICC and m_1 is the number of patients per therapist in the condition involving therapists. If $\rho_1 = .15$ and $m_1 = 10$, R equals 1.53. Power will be maximized for a given total sample size (N_T) if N_1 is 1.53 times the size of N_2. N_1 and N_2 can be calculated from N_T and R:

$$N_1 = \frac{N_T R}{R+1} \qquad (5)$$

$$N_2 = \frac{N_T}{R+1} \qquad (6)$$

Example power analysis: treatment versus treatment

Suppose we want to design an effectiveness study to compare cognitive therapy (CT) for depression versus treatment as usual (TAU). We plan to randomly assign patients to receive 16 sessions of either CT or TAU. A primary outcome measure will be the BDI, which the patients will complete prior to treatment and immediately following the 16th session. We would like to determine the detectable difference for the BDI, assuming 80% power, a two-tailed test, a Type I error rate of 5%, and anticipated sample size.

We will estimate treatment effects with an analysis of posttest BDI data adjusted for the baseline value of the BDI. We would like to recruit 200 total participants and would like to recruit 10 therapists for each condition ($k_1 = k_2 = 10$). Thus, each therapist will see 10 patients

($m_1 = m_2 = 10$), making the sample size for each condition 100 ($N_1 = N_2 = 100$). Given the analysis plan, we use the aggregate ICC estimate for the BDI from effectiveness studies calculated previously ($\bar{\rho}^I = .05$). We assume $\bar{\rho}^I$ will be equivalent in the CT and TAU conditions ($\rho_1 = \rho_2$). Because we are using a standardized metric, in this case Cohen's *d*, $\sigma_y^2 = 1$. The values for the *t* variates are taken from the *t* distribution with 10+10 − 2 = 18 *df*.

To determine the detectable difference, we insert the relevant values into Equation 2:

$$.505 = \sqrt{1\left(\frac{(1+(10-1).05)}{100}+\frac{(1+(10-1).($$

Thus, with 10 therapists per condition, each seeing 10 patients, and assuming $\rho_1 = \rho_2 = 0.05$, we would have 80% power to detect a treatment effect of $d = 0.505$. Given that the hypothetical study compares two active treatments, we would likely want to detect a smaller effect size. We can easily vary the values in Equation 2 to reflect the assumptions we make about the data. For example, suppose we can increase our total sample size to 260 patients by including more therapists per condition, more patients per therapist, or some combination. The first five columns of Table 3 illustrate the effects of manipulating either the number of therapists per condition or patients per therapist to bring the total sample size to 260. As can be seen in Table 3, increasing the number of therapists per condition has a greater effect on the detectable difference than increasing the number of patients per therapist. Likewise, power is increased if the number of patients per therapist is balanced across conditions. Both points are quite general and well established in the literature on group-randomized trials (Donner & Klar, 2000; Murray, 1998).

Table 3 Detectable difference varying the number of therapists per condition, patients per therapist, and intraclass correlation

				ρ_1 = .05	**ρ_1 = .10**	**ρ_1 = .15**	**ρ_1 = .20**
k $_1$	***k*** $_2$	***m*** $_1$	***m*** $_2$	**ρ_2 = .05**	**ρ_2 = .01**	**ρ_2 = .01**	**ρ_2 = .01**
13	13	10	10	.436	.443	.475	.505
10	10	16	10	.473	.483	.523	.561
10	10	14	12	.466	.475	.516	.554
10	10	13	13	.465	.474	.515	.552
10	10	12	14	.466	.474	.515	.552
10	10	10	16	.473	.479	.519	.556

Note. The power analyses assume 80% power, 5% Type I error rate, and two-tailed tests. The detectable difference is in the standardized mean difference metric (*d*). k_1 and k_2, number of therapists in Conditions 1 and 2, respectively; m_1 and m_2, number of patients per therapists in Conditions 1 and 2, respectively; ρ_1 and ρ_2, intraclass correlation for Conditions 1 and 2, respectively.

Up to this point, we assumed that ρ_1 and ρ_2 were equivalent. However, it is possible for ρ_1 and ρ_2 to differ systematically. Columns 6 through 8 of Table 3 provide detectable difference values, where ρ_1 and ρ_2 differ for the increased sample size of 260. For a given value of ρ_2, the detectable difference will increase as ρ_1 increases. As before, power is maximized by increasing the number of therapists per condition and balancing the patients per therapist.

Example power analysis: treatment versus comparison condition

Consider another effectiveness study in which we evaluate the effects of CT (Condition 1) for depression versus bibliotherapy (Condition 2) and use a design similar to that for the study described previously: random assignment to conditions, two time points, 16 weeks of treatment, and the BDI as a primary outcome variable. As before, we will estimate treatment effects with an analysis of posttest BDI data adjusted for the baseline value of the BDI and will use the aggregate ICC estimate for the BDI effectiveness studies ($\bar{\rho}^{I}$ = .05). We would like to recruit approximately 150 patients (N_T = 150) and would like the therapists in the CT condition to treat 10 patients each (m_1 = 10). The optimal allocation ratio (*R*) is:

$$1.20 = \sqrt{(1 + (10 - 1)0.05)}$$

so that the CT condition should have 1.20 times as many patients as the bibliotherapy condition.

Using *R* and N_T, we calculate N_1 and N_2 as follows:

$$N_1 = 81.82 = \frac{150*1.20}{1.20 + 1}$$

$$N_2 = 68.18 = \frac{150}{1.20 + 1}$$

N_1 will need to be rounded to 90 as that will provide an integer value for the number of therapists in the CT condition. Because there are no therapists in the bibliotherapy condition, we round N_2 to 75 to maintain an allocation ratio of 1.20. The values for the *t* variates are taken from the *t* distribution with 9+75 − 2 = 82 *df*.

We insert the relevant values into Equation 3 to determine the detectable difference:

$$.487 = \sqrt{1\left(\frac{(1 + (10 - 1)0.05)}{90} + \frac{1}{75}\right)(1.99 + .}$$

Thus, with a total sample of 165 patients, with nine therapists each seeing 10 patients in the CT condition and 75 patients in the bibliotherapy condition, we would have 80% power to detect a treatment effect of $d = 0.487$. As before, we can vary the parameters in Equation 3 as needed.

Discussion

Estimated intraclass correlations

The distributions for $\hat{\rho}_{pre}$, $\hat{\rho}_{post}$, and $\hat{\rho}'_{post}$ as calculated in this study were symmetric and quite similar. The 25th, 50th, and 75th percentiles were − .099, − .014, and .063 for $\hat{\rho}_{pre}$; − .113, − .026, and .046 for $\hat{\rho}_{post}$; and − .104, − .018, and .079 for $\hat{\rho}'_{post}$. Importantly, 54.6%, 63.7%, and 57% of the estimates for $\hat{\rho}_{pre}$, $\hat{\rho}_{post}$, and $\hat{\rho}'_{post}$ were negative. Noting that all previously published estimates were positive with an average value of about .08, we conducted a series of post hoc investigations help us understand the apparent discrepancy.

Inspection of the published ICCs in psychotherapy research revealed that those investigators likely did not allow negative values (e.g. Crits-Christoph et al., 1991). This practice is so common that it has a name (the nonnegativity constraint; Swallow & Monahan, 1984) and is the default in many software packages used to estimate components of variance (e.g. SAS PROC MIXED, HLM). This practice reflects the common interpretation of an ICC as the proportion of variance, which cannot be negative; as a result, many researchers fix negative ICC estimates to zero (cf. Maxwell & Delaney, 2004). At the same time, this practice ignores the more general meaning of an ICC as a correlation, which can be negative (Kenny et al., 2002, p. 127; Pinheiro & Bates, 2000, p. 228; Snedecor & Cochran, 1989, p. 243; Kish, 1965, p. 163). We hypothesized that the apparent discrepancy between the published ICCs and our own findings had do to the nonnegativity constraint in the calculation of the ICCs in the published studies.

The theoretical range of ICCs is $-1/(m-1)$ to 1 (Kenny et al., 2002; Pinheiro & Bates, 2000; Snedecor & Cochran, 1989; Kish,1965, p. 163). Thus, if each therapist sees two patients, ICCs can range from − 1 to 1. If each therapist sees three patients, ICCs can range from − .5 to 1. If each therapist sees 10 patients, ICCs can range from − .11 to 1. As m approaches infinity, the lower bound of the ICC approaches zero (see Figure 1). Given that m in the studies examined for this study ranged from 2.2 to 51.1, ICCs could range from − .83 to 1 in the smallest study and from − .02 to 1 in the largest study; the observed range was − .475 to .579.

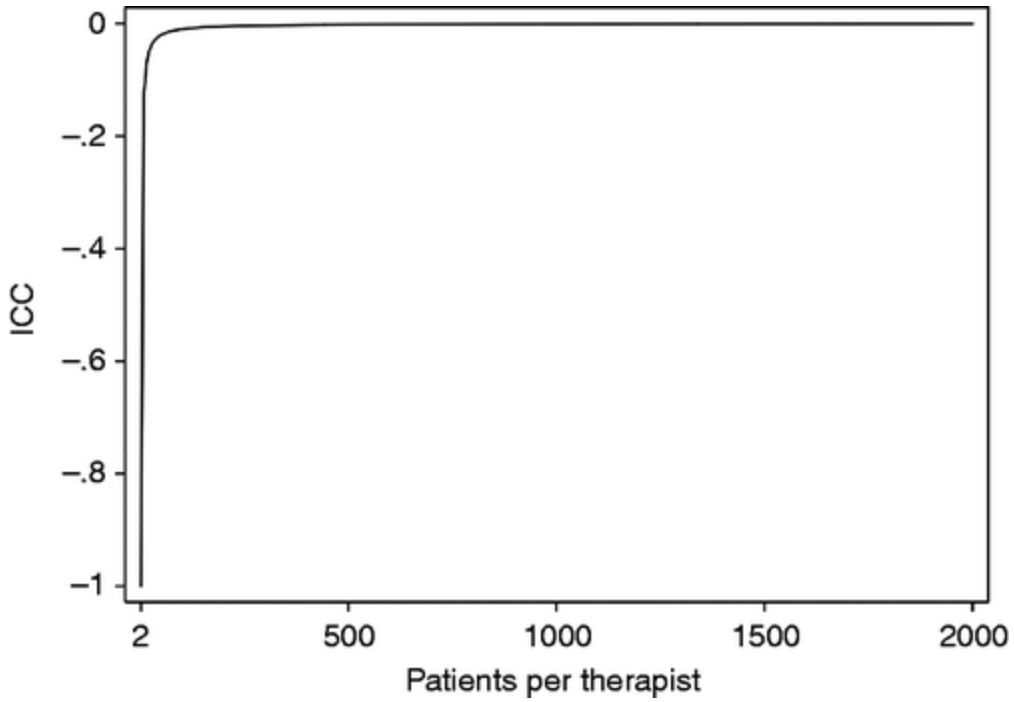

Figure 1 The relationship between the number of patients per therapist and the lower bound of an intraclass correlation (ICC).

We used Monte Carlo simulation to investigate the impact of *m* on the frequency of negative ICC estimates. We also varied *k*(number of therapists) and the population ICC. We allowed *m* to equal 2, 4, 8, 32, 64, 128, or 256; *k* to equal 5, 10, 20, or 40; and the ICC to equal − .001, .001, .05, and .1. We generated 1,500 data sets for each cell in the simulation. For each replication we estimated a one-way ANOVA and used Equation 1 to calculate the ICC. All data were generated and analyzed in Stata (version 11; StataCorp, 2009).

Table 4 presents the percentage of negative estimates by *m*, *k*, and population ICC. Three patterns emerge from the results. First, when *m* and *k* were relatively small, many of the estimates were negative. When *k* = 5 and *m* = 4, similar to the median values in our sample of studies, the percentage of negative ICCs ranged from 40 to 57%. Second, as cluster size increased, the number of negative ICCs declined except when the population ICC was − .001. In that case, increasing *k* and *m* increased the percentage of negative estimates. Third, *k* had less influence on the proportion of negative estimates than *m*, especially when *m* was small. Together, these results suggest that our observed ICCs are very plausible given the sample sizes in our sample of studies and common in psychotherapy research (see also Figure 2).

Table 4 Percentage of negative intraclass correlation estimates in simulated data

Intraclass correlation coefficient				
	− .001	**.001**	**.05**	**.10**

m	*k* = 5	*k* = 10	*k* = 20	*k* = 40	*k* = 5	*k* = 10	*k* = 20	*k* = 40	*k* = 5	*k* = 10	*k* = 20	*k* = 40	*k* = 5	*k* = 10	*k* = 20	*k* = 40
2	51	50	52	49	51	50	52	48	46	45	44	36	43	39	36	26
4	57	56	55	53	57	55	54	52	48	43	33	25	40	32	20	9
8	61	58	55	54	60	57	53	51	44	29	19	9	31	15	6	1
16	61	57	56	55	59	54	52	50	29	17	6	1	16	5	1	0
32	62	60	57	59	58	55	50	48	18	6	1	0	7	1	0	0
64	63	62	62	63	56	51	48	41	8	1	0	0	3	0	0	0
128	66	67	69	74	52	47	40	31	3	0	0	0	1	0	0	0
256	76	80	84	92	48	38	29	18	1	0	0	0	0	0	0	0

Note. m, patients per therapist; k, therapists.

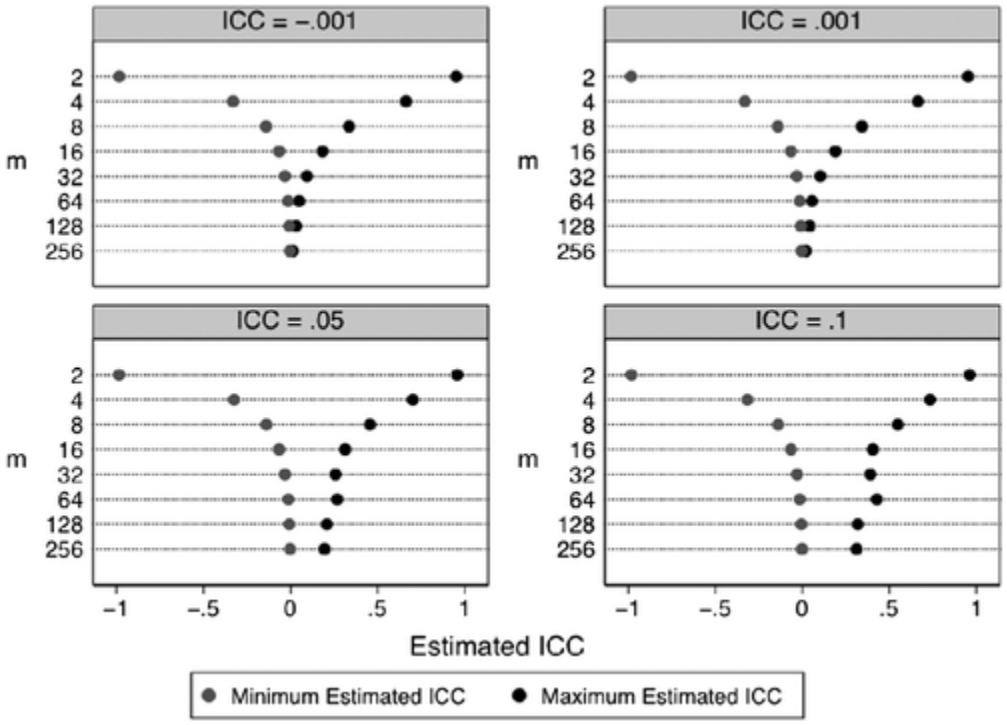

Figure 2 Simulated minimum and maximum values of the intraclass correlation (ICC) stratified by patients per therapist (*m*) and population ICC. Each cell of the simulation was replicated 1500 times. The number of therapists (*k*) was five. The pattern of results with other values of *k* was similar. Full results are available from Scott A. Baldwin.

Fixing negative values to zero creates an upward bias in the ICC estimate. To test our post hoc hypothesis about the nonnegativity constraint, we calculated the bias that would occur by fixing negative values to zero in our simulated data. We calculated ICCs twice: the first allowing negative estimates and the second fixing the negative estimates to zero. We calculated the bias in each case by subtracting the population ICC from the mean estimated ICC across the 1,500 replications for each cell. A positive result indicates that the mean ICC overestimated the population ICC, whereas a negative result indicates that the mean ICC underestimated the population ICC. Figure 3 displays the magnitude of bias across levels of *m* and the population ICC. Figure 3 only presents results for $k = 5$, although the pattern of results was the same across all levels of *k*. When we allowed for negative estimates, there was a very slight negative bias when *m* was very small, especially for smaller ICCs, and no bias when *m* was larger. When we fixed negative estimates to zero, there was a much larger positive bias when *m* was small, especially for smaller ICCs, and no bias when *m* was larger.

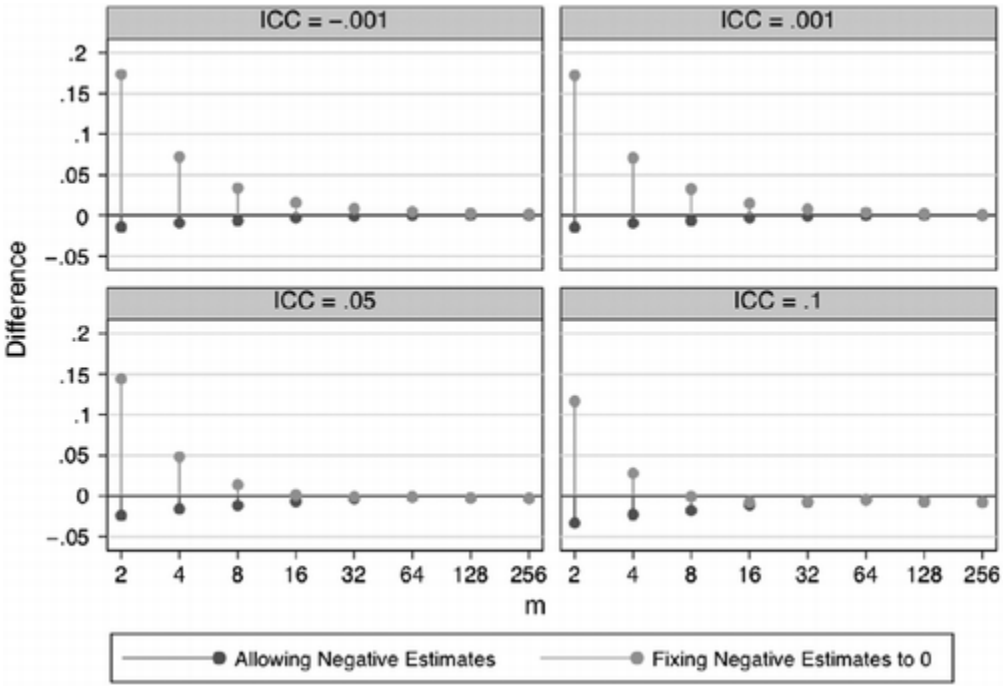

Figure 3 Difference between the average estimated intraclass correlation (ICC) and the population ICC stratified by patients per therapist (*m*) and population ICC. Each cell of the simulation was replicated 1500 times. The number of therapists (*k*) was five. The pattern of results with other values of *k* was similar. Full results are available from Scott A. Baldwin.

Recall that a meta-analysis of ICCs averaged about .08 based on a median *m* of about 7. If the population ICC in those studies was really − .001, Figure 3 suggests that fixing negative estimates to zero would result in a bias of about .04, for an estimated average ICC of .039. If the population ICC in those studies was really .05, Figure 3 suggests that fixing negative estimates to zero would result in a bias of about .025, for an estimated average ICC of .075. Thus, it is quite

likely that if the published estimates used the nonnegativity constraint, the true average ICC is between .001 and .05. This would largely explain the apparent discrepancy between our findings and the previously published results. Taken together, the results of our simulations provide considerable support for the validity of the estimates we report here.

Implications for future research

Statistical dependencies within therapists are an important methodological issue that affects interpretation of intervention trials. This is true whether the dependencies are preexisting, are due to the therapist, or have some other origin. To address this issue, researchers must account for dependencies when planning their studies and when analyzing their data. To plan studies adequately, researchers need ICC estimates; they also need to know how to select and use those estimates. Our primary aims were (1) to report an initial database of ICC estimates associated with therapist for a variety of measures and treatment conditions from a sample of recently published studies in leading psychotherapy journals and (2) to provide guidelines for the selection and use of ICC estimates for power calculations.

This study has made clear that the size of typical psychotherapy studies leads to imprecise ICC estimates, especially compared with other disciplines. For example, the ICC estimates for academic achievement reported in Hedges & Hedberg (2007) were based on hundreds and sometimes thousands of schools and thousands of students. In the current database, the median number of therapists contributing to an ICC was five and the median number of patients per therapist was five. Although these values can be reasonably increased in future studies, most psychotherapy studies typically have relatively small numbers of therapists (< 20), each treating a modest number of patients (10–20). Thus, ICC estimates in psychotherapy research will rarely match the precision of ICC estimates in other disciplines. As a result, we recommend that researchers use a range of estimates. This will become less of an issue as the ICC database gets larger because researchers will be able to meta-analytically combine estimates from many studies. Therefore, it is critical for researchers to contribute to the ICC database by routinely reporting ICC estimates.

The results of the illustrative power analyses underscore three points. First, when the ICC is greater than zero, the power to detect a treatment effect is reduced compared with when the ICC is zero (or negative). Consequently, planning for statistical dependence when designing studies is essential. If researchers do not attend to these issues until the analysis stage of their study, a proper analysis will be underpowered to detect a treatment effect despite having an otherwise well-designed study. Second, when comparing a treatment involving therapists with a comparison condition that does not, power will be maximized by allocating more patients to the treatment condition. As Roberts & Roberts (2005) pointed out, if the treatment is more effective than the control, this unequal allocation has the practical benefit of providing the treatment to more people. Third, reducing the ICC via standardization (Crits-Christoph, Tu, & Gallop, 2003) or covariates (Murray & Blitstein, 2003) may produce the largest increase in power at the lowest

cost, followed by increasing the number of therapists and then increasing the number of patients per therapist.

The aggregate ICC estimates were not statistically significant. Consequently, investigators might be tempted to conclude that the ICCs can be ignored or that we should conduct an initial significance test for the ICC and only model therapists as a random effect if it is significant (cf. Crits-Christoph et al., 2003). We disagree strongly for three reasons. First, the power to detect a significant ICC is typically too low to allow such tests to be trustworthy, even with liberal *p* values or in a meta-analysis (Kenny, Kashy, & Bolger, 1998; Murray, 1998; Roberts & Roberts, 2005). For example, if a study included five therapists per condition with 10 patients per therapist, the study would have only 23% power to detect an ICC of .05. A study with 10 therapists per condition with 10 patients per therapist would have only 37% power to detect an ICC of .05 (Winer, Brown, & Michels, 1991). Thus, even if a study is adequately powered to detect a desired treatment effect when the population ICC is .05, it may include too few therapists to be adequately powered to detect an ICC of that size. Trustworthy significance tests in meta-analysis will not likely be available until researchers consistently report ICCs. Second, even if ICCs are *estimated* as zero or close to zero, the degrees of freedom still need to be based on the number of therapists, not the number of patients (Baldwin et al., 2005; Murray et al., 1996; Pals et al.,2008). Third, the problems created by statistical dependence do not depend on the statistical significance of the ICC estimate (Kenny et al., 2002; Murray, 1998; Roberts & Roberts, 2005), and instead are a function of both the magnitude of the ICC and the number of patients treated by each therapist. Consequently, we join methodologists in public health and psychology and recommend that researchers model the dependencies in their data regardless of the statistical significance of the ICC to safeguard the Type I error rate in their studies (Donner & Klar, 2000; Kenny et al., 2002; Murray, 1998).

Conclusions

Accounting for statistical dependencies associated with therapist has proven to be a significant challenge in psychotherapy research. A major hurdle has been that accounting for statistical dependencies increases the cost and complexity of the already expensive and difficult process of psychotherapy research. Indeed, accounting for statistical dependencies associated with therapist involves not only recruiting more patients but recruiting, training, and supervising more therapists. Nevertheless, accounting for statistical dependencies is a priority because ignoring them threatens the validity of the inferences drawn about treatment efficacy. Several disciplines face similar methodological challenges, such as education and public health. Researchers in these disciplines have begun to adapt their research design and analytic methods to address these issues (Varnell, Murray, Janega, & Blitstein, 2004; Murray, Pals, Blitstein, Alfano, & Lehman, 2008). We hope that the material presented in this report will help psychotherapy researchers move in the same direction.

Acknowledgement

This research was supported by National Institutes of Health Research Grant MH73203-01.

Notes

1. A table listing all ICCs is available from Scott A. Baldwin or can be downloaded at http://psychology.byu.edu/Faculty/SBaldwin/Home.dhtml.

References

References marked with an asterisk indicate studies contributing ICCs.

1. *Abramowitz, J. S., Foa, E. B. and Franklin, M. E. 2003. Exposure and ritual prevention for obsessive-compulsive disorder: Effects of intensive versus twice-weekly sessions. *Journal of Consulting and Clinical Psychology*, 71: 394–398.

2. Baldwin, S. A., Murray, D. M. and Shadish, W. R. 2005. Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology*, 73: 924–935.

3. *Baldwin, S. A., Wampold, B. E. and Imel, Z. E. 2007. Untangling the alliance–outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology*, 75: 842–852.

4. Beck, A. T., Steer, R. A. and Brown, G. K. 1996. *Manual for Beck Depression Inventory–II*, San Antonio, TX: Psychological Corporation.

5. Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. and Erbaugh, J. 1961. An inventory for measuring depression. *Archives of General Psychiatry*, 4: 561–571.

6. Bergin, A. E. 1966. Some implications of psychotherapy research for therapeutic practice. *Journal of Abnormal Psychology*, 71: 235–246.

7. Blitstein, J. L., Hannan, P. J., Murray, D. M. and Shadish, W. R. 2005. Increasing the degrees of freedom in existing group randomized trials: The df* approach. *Evaluation Review*, 29: 241–267.

8. *Carlbring, P., Bohman, S., Brunt, S., Buhrman, M., Westling, B. E., Ekselius, L. and Andersson, G. 2006. Remote treatment of panic disorder: A randomized trial of Internet-based cognitive behavior therapy supplemented with telephone calls. *American Journal of Psychiatry*, 163: 2119–2125.

9. *Carroll, K. M., Fenton, L. R., Ball, S. A., Nich, C., Frankforter, T. L., Shi, J. and Rounsaville, B. J. 2004. Efficacy of disulfiram and cognitive behavior therapy in cocaine-dependent outpatients: A randomized placebo-controlled trial. *Archives of General Psychiatry*, 61: 264–272.

10. *Christensen, A., Atkins, D. C., Berns, S., Wheeler, J., Baucom, D. H. and Simpson, L. E. 2004. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *Journal of Consulting and Clinical Psychology*, 72: 176–191.

11. Cornfield, J. 1978. Randomization by group: A formal analysis. *American Journal of Epidemiology*, 108: 100–102.

12. Crits-Christoph, P., Baranackie, K., Kurcias, J. S., Beck, A. T., Carroll, K., Perry, K. and Zitrin, C. 1991. Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research*, 1: 81–91.

13. Crits-Christoph, P. and Mintz, J. 1991. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology*, 59: 20–26.

14. Crits-Christoph, P., Tu, X. and Gallop, R. 2003. Therapists as fixed versus random effects—Some statistical and conceptual issues: A comment on Siemer and Joormann (2003). *Psychological Methods*, 8: 518–523.

15. *Dinger, U., Strack, M., Leichsenring, F., Wilmers, F. and Schauenburg, H. 2008. Therapist effects on outcome and alliance in inpatient psychotherapy. *Journal of Clinical Psychology*, 64: 344–354.

16. Donner, A. 1986. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistics Review*, 54: 67–82.

17. Donner, A. and Klar, N. 2000. *Design and analysis of cluster randomization trials in health research*, London: Arnold.

18. *Ehlers, A., Clark, D. M., Hackmann, A., McManus, F., Fennell, M., Herbert, C. and Mayou, R. 2003. A randomized controlled trial of cognitive therapy, a self-help booklet, and repeated assessments as early interventions for posttraumatic stress disorder.*Archives of General Psychiatry*, 60: 1024–1032.

19. Elkin, I., Falconnier, L., Martinovich, Z. and Mahoney, C. 2006. Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Psychotherapy Research*, 16: 144–160.

20. Fleiss, J. L., Levin, B. and Paik, M. C. 2003. *Statistical methods for rates and proportions*, 3rd ed, Hoboken, NJ: Wiley.

21. Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B. and Pee, D. 1996. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15: 1069–1092.

22. Gulliford, M. C., Ukomunne, O. C. and Chinn, S. 1999. Components of variance and intraclass correlation for the design of community-based surveys and intervention studies. *American Journal of Epidemiology*, 149: 876–883.

23. Hedges, L. V. and Hedberg, E. C. 2007. Intraclass correlation values for planning group-randomized trials in education.*Educational Evaluation and Policy Analysis*, 29: 60–87.

24. Higgins, J. P. T. and Thompson, S. G. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21: 1539–1558.

25. Kenny, D. A., Kashy, D. A. and Bolger, N. 1998. "Data analysis in social psychology". In *The handbook of social psychology*, Edited by: Gilbert, D. T., Fiske, S. T. and Lindzey, G. Vol. 1, 233–265. New York, NY: Oxford University Press.

26. Kenny, D. A., Mannetti, L., Peirro, A., Livi, S. and Kashy, D. A. 2002. The statistical analysis of data from small groups.*Journal of Personality and Social Psychology*, 83: 126–137.

27. *Kim, D. -M., Wampold, B. E. and Bolt, D. M. 2006. Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research*, 16: 161–172.

28. Kish, L. 1965. *Survey sampling*, New York, NY: Wiley.

29. *Koch, E. I., Spates, C. R. and Himle, J. A. 2004. Comparison of behavioral and cognitive-behavioral one-session exposure treatments for small animal phobias. *Behaviour Research and Therapy*, 42: 1483–1504.

30. *Kuyken, W. 2004. Cognitive therapy outcome: The effects of hopelessness in a naturalistic outcome study. *Behaviour Research and Therapy*, 42: 631–646.

31. *Lange, A., Rietdijk, D., Hudcovicova, M., van de Ven, J. P., Schrieken, B. and Emmelkamp, P. M. 2003. Interapy: A controlled randomized trial of the standardized treatment of posttraumatic stress through the Internet. *Journal of Consulting and Clinical Psychology*, 71: 901–909.

32. *Lincoln, T. M., Rief, W., Hahlweg, K., Frank, M., von Witzleben, I., Schroeder, B. and Fiegenbaum, W. 2003. Effectiveness of an empirically supported treatment for social phobia in the field. *Behaviour Research and Therapy*, 41: 1251–1269.

33. Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S. and Stiles, W. B. 2007. Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology*, 54: 32–39.

34. *Marijuana Treatment Project Research Group. 2004. Brief treatments for cannabis dependence: Findings from a randomized multisite trial. *Journal of Consulting and Clinical Psychology*, 72: 455–466.

35. Martindale, C. 1978. The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 46: 1526–1530.

36. Maxwell, S. E. and Delaney, H. D. 2004. *Designing experiments and analyzing data: A model comparison approach*, 2nd ed., Mahwah, NJ: Erlbaum.

37. *Merrill, K. A., Tolbert, V. E. and Wade, W. A. 2003. Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology*, 71: 404–409.

38. Murray, D. M. 1998. *Design and analysis of group-randomized trials*, New York, NY: Oxford University Press.

39. Murray, D. M. and Blitstein, J. L. 2003. Methods to reduce the impact of intraclass correlation in group-randomized trials.*Evaluation Review*, 27: 79–103.

40. Murray, D. M., Hannan, P. J. and Baker, W. L. 1996. A Monte Carlo study of alternative responses to intraclass correlation in community trials. Is it ever possible to avoid Cornfield's penalties?. *Evaluation Review*, 20: 313–337.

41. Murray, D. M., Pals, S. P., Blitstein, J. L., Alfano, C. M. and Lehman, J. 2008. Design and analysis of group-randomized trials in cancer: A review of current practices. *Journal of the National Cancer Institute*, 100: 483–491.

42. Murray, D. M., Rooney, B. L., Hannan, P. J., Peterson, A. V., Ary, D. V., Biglan, A. and Schinke, S.P. 1994. Intraclass correlation among common measures of adolescent smoking: Estimates, correlates, and applications in smoking prevention studies. *American Journal of Epidemiology*, 140: 1038–1050.

43. Murray, D. M., Varnell, S. P. and Blitstein, J. L. 2004. Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, 94: 423–432.

44. Okiishi, J. C., Lambert, M. J., Nielsen, L. and Ogles, B. M. 2003. Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology and Psychotherapy*, 10: 361–373.

45. Pals, S. P., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J. and Baker, W. L. 2008. Individually randomized group treatment studies: Are the most frequently used analytic models misleading?. *American Journal of Public Health*, 98: 1418–1424.

46. Pinheiro, J. C. and Bates, D. M. 2000. *Mixed-effects models in S and S-Plus*, New York, NY: Springer.

47. Roberts, C. 1999. The implications of variation in outcome between health professionals for the design and analysis of randomized controlled trials. *Statistics in Medicine*, 18: 2605–2615.

48. Roberts, C. and Roberts, S. A. 2005. Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, 2: 152–162.

49. Satterthwaite, F. W. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2: 110–140.

50. Schnurr, P. P., Friedman, M. J., Lavori, P. W. and Hsieh, F. Y. 2001. Design of Department of Veterans Affairs Cooperative Study No. 420: Group treatment of posttraumatic stress disorder. *Controlled Clinical Trials*, 22: 74–88.

51. Snedecor, G. W. and Cochran, W. G. 1989. *Statistical methods*, 8th ed., Ames, IA: Iowa State University Press.

52. StataCorp (2009). Stata Statistical Software (Version 11) [Computer software]. College Station, TX: Author

53. Swallow, W. H. and Monahan, J. F. 1984. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, 26: 47–57.

54. *Szapocznik, J., Feaster, D. J., Mitrani, V. B., Prado, G., Smith, L., Robinson-Batista, C. and Robbins, M. S. 2004. Structural ecosystems therapy for HIV-seropositive African American women: Effects on psychological distress, family hassles, and family support. *Journal of Consulting and Clinical Psychology*, 72: 288–303.

55. *Taylor, S., Thordarson, D. S., Maxfield, L., Fedoroff, I. C., Lovell, K. and Ogrodniczuk, J. 2003. Comparative efficacy, speed, and adverse effects of three PTSD treatments: Exposure therapy, EMDR, and relaxation training. *Journal of Consulting and Clinical Psychology*, 71: 330–338.

56. *Trepka, C., Rees, A., Shapiro, D. A., Hardy, G. E. and Barkham, M. 2004. Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy and Research*, 28: 143–157.

57. *van Minnen, A., Hoogduin, K. A., Keijsers, G. P., Hellenbrand, I. and Hendriks, G. J. 2003. Treatment of trichotillomania with behavioral therapy or fluoxetine: A randomized, waiting-list controlled study. *Archives of General Psychiatry*, 60: 517–522.

58. Varnell, S., Murray, D. M., Janega, J. B. and Blitstein, J. L. 2004. Design and analysis of group-randomized trials: A review of recent practices. *American Journal of Public Health*, 94: 393–399.

59. Verma, V. and Le, T. 1996. An analysis of sampling errors for the demographic and health surveys. *International Statistical Review*, 64: 265–294.

60. Wampold, B. E. 2001. *The great psychotherapy debate*, Mahwah, NJ: Erlbaum.

61. *Wampold, B. E. and Brown, G. S. 2005. Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology*, 73: 914–923.

62. Wampold, B. E. and Serlin, R. C. 2000. The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods*, 5: 425–433.

63. *Watson, J. C., Gordon, L. B., Stermac, L., Kalogerakos, F. and Steckley, P. 2003. Comparing the effectiveness of process-experiential with cognitive-behavioral psychotherapy in the treatment of depression. *Journal of Consulting and Clinical Psychology*, 71: 773–781.

64. Winer, B. J., Brown, D. R. and Michels, K. 1991. *Statistical principles in experimental design*, New York, NY: McGraw-Hill.

65. Zucker, D. M. 1990. An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement*, 50: 731–738.